

KDD E MINERAÇÃO DE DADOS

Etapas do Processo de KDD

Prof. Ronaldo R. Goldschmidt

ronaldo@de9.ime.eb.br

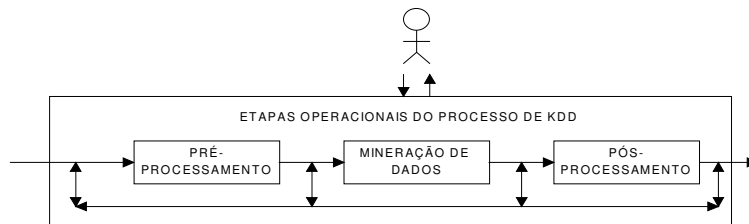
rribeiro@univercidade.br

geocities.yahoo.com.br/ronaldo_goldschmidt

ETAPAS DO PROCESSO DE KDD

Processo de Descoberta do Conhecimento em Bases de Dados

- Visão Pragmática [Goldschmidt et al., 2002a]:



- Operações e Métodos de KDD

ETAPAS DO PROCESSO DE KDD

Pré-Processamento dos Dados

Compreende as ações para adequar os dados aos algoritmos de mineração de dados a serem aplicados.

Tipos de Variáveis/Atributos

- Nominais ou Categóricas
- Discretas
- Contínuas

ETAPAS DO PROCESSO DE KDD

Tipos de Variáveis/Atributos: Nominais ou Categóricas

- Utilizadas para nomear ou atribuir rótulos a objetos.
- Valores: conjunto finito e pequeno de estados.
- Não há ordem entre os valores.
- Podem ser representados por tipos de dados alfanuméricos.

Exemplo:

Estado Civil:
Solteiro, Casado, Viúvo, Divorciado, etc...

ETAPAS DO PROCESSO DE KDD

Tipos de Variáveis/Atributos: Discretas

- Similar às variáveis nominais.
- Há ordem (com significado) entre os valores.
- Podem ser representados por tipos alfanuméricos.

Exemplos:

Dia da Semana: Segunda-Feira, Terça-Feira, Quarta-Feira, etc...,
Faixa Etária,
Faixa de Renda, etc...

ETAPAS DO PROCESSO DE KDD

Tipos de Variáveis/Atributos: Contínuas

- Variáveis Quantitativas.
- Valores: conjuntos finitos ou infinitos, ordenados.
- Tipos de dados numéricos.

Exemplos:

Renda, Idade, Altura, Etc...

ETAPAS DO PROCESSO DE KDD

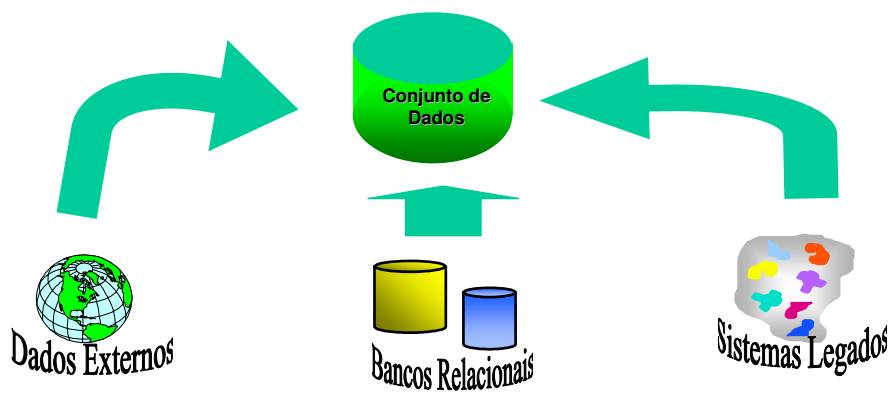
Exemplos de Operações de Pré-Processamento

- Seleção/Redução de Dados
- Limpeza
- Codificação
- Normalização de Dados
- Enriquecimento
- Construção de Atributos
- Correção de Prevalência
- Partição dos Dados

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

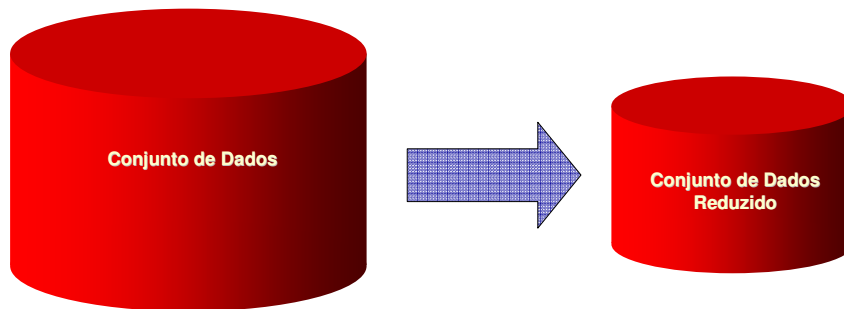
Definição 1: Extração dos dados de diversas fontes e carga no conjunto de dados a ser analisado.



ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

Definição 2: Escolha dentre os dados de um conjunto de dados, quais deverão ser efetivamente considerados na análise.



ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

Definição 3: Ocorre em dois momentos:

1°. Extração dos dados de diversas fontes e carga no conjunto de dados a ser analisado

2°. Escolha dentre os dados de um conjunto de dados, quais deverão ser efetivamente considerados na análise

Combinação das definições anteriores

Bastante comum na prática

Cópia do Conjunto de Dados em quaisquer das definições

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

A cópia dos dados em um conjunto separado dos dados transacionais → Evita interferência tanto no cotidiano quanto nas ações de mineração.

Consideremos que os dados estejam organizados em um mesmo conjunto: estrutura tabular bidimensional.

Assim, adotaremos a definição 2 para seleção de dados:

Definição 2: Escolha dentre os dados de um conjunto de dados, quais deverão ser efetivamente considerados na análise.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal:** escolha de casos
 - Segmentação do Banco de Dados

Exemplo: Analisar somente clientes com residência própria

```
SELECT *  
FROM CLIENTE  
WHERE TP_RES="P";
```

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
 - **Eliminação Direta de Casos (variação da anterior)**

Exemplo: Analisar somente clientes com residência própria

```
DELETE FROM CLIENTE  
WHERE TP_RES<>'P';
```

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
 - **Amostragem Aleatória**

Há várias abordagens. Sejam N o total de registros e n o número de amostras desejadas ($n < N$).

- **Amostragem Aleatória Simples sem Reposição**
 - **Probabilidade de seleção: $1/N$, que aumenta ao longo do processo.**
 - **Todos os registros selecionados são excluídos do conjunto de dados original para evitar nova seleção.**

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
 - **Amostragem Aleatória**

Há várias abordagens. Sejam N o total de registros e n o número de amostras desejadas ($n < N$).

- **Amostragem Aleatória Simples com Reposição**
 - **Probabilidade de seleção: $1/N$.**
 - **Todos os registros selecionados são mantidos no conjunto de dados original podendo sofrer nova seleção.**

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
 - **Amostragem Aleatória**

Há várias abordagens. Sejam N o total de registros e n o número de amostras desejadas ($n < N$).

- **Amostragem de Clusters**
 - **Registros devem estar agrupados em M clusters (grupos).**
 - **Cada repetição do processo de amostragem pressupõe a escolha do cluster.**

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
 - **Amostragem Aleatória**

Há várias abordagens. Sejam N o total de registros e n o número de amostras desejadas ($n < N$).

- **Amostragem Estratificada**
 - Registros devem estar separados em grupos distintos, previamente definidos segundo algum critério.
 - O processo de amostragem deve ocorrer dentro de cada grupo.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Horizontal: escolha de casos**
 - **Agregação de Informações**
- Dados em um maior nível de detalhe são agrupados em um nível mais consolidado.

Exemplo: Somar compras por cliente por período.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Vertical: escolha de características/atributos**

Sendo S um conjunto de dados com atributos $A_1, A_2, A_3, \dots, A_n$, o problema da **redução de dados vertical** consiste em **identificar** qual das 2^n combinações entre tais **atributos** deve ser considerada no processo de descoberta de conhecimento.

Tem como **objetivo** procurar **encontrar** um **conjunto mínimo de atributos** de tal forma que a **informação original** seja ao **máximo preservada**.

Quanto maior o valor de n , maior o desafio na escolha dos atributos.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Vertical: escolha de características/atributos – Motivações:**

- ✓ Um conjunto de atributos bem selecionado pode conduzir a modelos de conhecimento mais concisos e com maior precisão;
- ✓ A eliminação de um atributo é muito mais significativa em termos de redução do tamanho de um conjunto de dados do que a exclusão de um registro.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Vertical: escolha de características/atributos – Motivações:**

- ✓ Se o método de seleção for rápido, o tempo de processamento necessário para utilizá-lo e, em seguida, aplicar o algoritmo de mineração de dados em um subconjunto dos atributos, pode ser inferior ao tempo de processamento para aplicar o algoritmo de mineração sobre todo o conjunto de atributos;

ETAPAS DO PROCESSO DE KDD

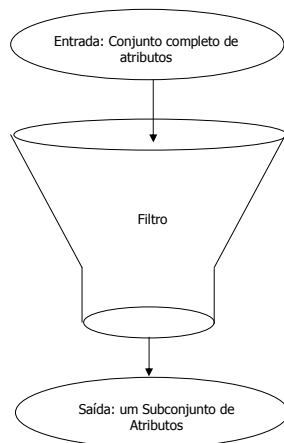
Seleção/Redução de Dados

- **Vertical: escolha de características/atributos – Abordagens segundo (Freitas, 2002):**

- ✓ Abordagem Independente de Modelo (*Filter*)
- ✓ Abordagem Dependente de Modelo (*Wrapper*)

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical – Abordagens

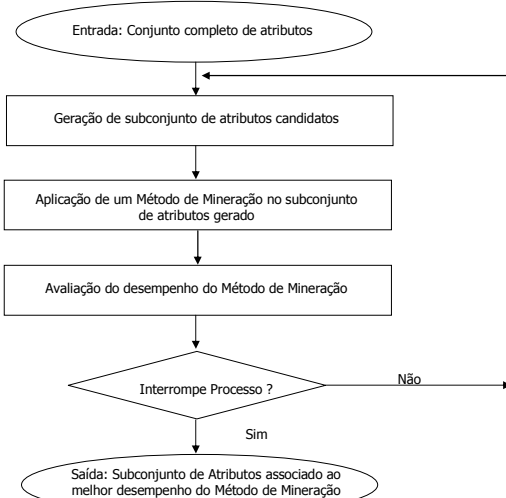


Abordagem Independente de Modelo (*Filter*):

A seleção de atributos é realizada sem considerar o algoritmo de mineração de dados que será aplicado.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical – Abordagens



Abordagem Dependente de Modelo (*Wrapper*): Consiste em experimentar o algoritmo de mineração de dados para cada conjunto de atributos e avaliar os resultados obtidos.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Vertical: escolha de características/atributos – Estratégias:**

- ✓ Seleção Sequencial para Frente (*Forward Selection*)

- ✓ Seleção Sequencial para Trás (*Backward Selection*)

- ✓ Combinação das Anteriores (*Mixed Selection*)

Podem ser utilizadas em quaisquer das abordagens citadas

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Vertical: escolha de características/atributos – Estratégias:**

- ✓ Seleção Sequencial para Frente (*Forward Selection*)

- Subconjunto de atributos candidatos começa vazio.
- O processo é iterativo.
- Cada atributo é adicionado ao subconjunto de candidatos
- O subconjunto de candidatos é avaliado segundo medida de qualidade.
- Ao final de cada iteração, é incluído no subconjunto de atributos candidatos, aquele atributo que tenha maximizado a medida de qualidade considerada

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Vertical: escolha de características/atributos – Estratégias:**

- ✓ Seleção Sequencial para Trás (*Backward Selection*)

- Contrário da Seleção Sequencial para Frente.
- Subconjunto de candidatos começa completo.
- Cada atributo é retirado do subconjunto, que é avaliado segundo alguma medida de qualidade.
- Ao final de cada iteração, é excluído do subconjunto de candidatos, aquele atributo que tenha minimizado a medida de qualidade.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Vertical: escolha de características/atributos – Estratégias:**

- ✓ Combinação Anteriores (*Mixed Selection*)

- A seleção para frente e a seleção para trás são combinadas.
- A cada iteração, o algoritmo seleciona o melhor atributo (incluindo-o no subconjunto de atributos candidatos) e remove o pior atributo dentre os remanescentes do conjunto de atributos.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados

- **Vertical: escolha de características/atributos – Observações:**

- ✓ Tarefa de Otimização – Solução adequada por Algoritmos Genéticos, C4.5, PSO, dentre outras metaheurísticas

- ✓ Exemplo de Medida de Qualidade – Taxa de Inconsistência

Sexo	Est_Civil	Result	Count(*)
M	C	A	2
M	C	I	1
F	S	A	3
M	S	I	1
F	C	I	2
M	V	A	1

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: Eliminação Direta de Atributos

Consiste da escolha dos atributos desejados, desprezando os demais.

Depende do conhecimento prévio sobre o problema.

Heurísticas Importantes para eliminação de atributos:

- a) Atributos com valores constantes → não adicionam informação.
- b) Atributos que sejam identificadores → particularizam objetos.

ETAPAS DO PROCESSO DE KDD

Seleção/Redução de Dados Vertical

Método: PCA – Análise de Componentes Principais

Técnica de baixo custo computacional aplicável a dados numéricos.

Consiste na transformação dos dados para um novo espaço com dimensão inferior ao original.

O novo espaço fica caracterizado por um novo conjunto de eixos, ortonormais entre si, ordenados em ordem decrescente de variância.

ETAPAS DO PROCESSO DE KDD

Redução de Valores Nominais

Método: Hierarquia entre Atributos

Especialista no domínio da aplicação especifica a hierarquia.

Exemplo: Atributo Endereço

Redução de valores distintos:

logradouro $\subset$ bairro $\subset$ cidade $\subset$ unidade da federação.

ETAPAS DO PROCESSO DE KDD

Redução de Valores Nominais

Método: Hierarquia entre Valores

Especialista no domínio da aplicação especifica possíveis abstrações entre os valores de cada atributo.

Exemplo: Atributo Produto_Vendido

Redução de valores distintos:

tênis $\subset$ calçado, sapato $\subset$ calçado, sandália $\subset$ calçado,
bermuda $\subset$ roupa, calça $\subset$ roupa, camisa $\subset$ roupa e paletó $\subset$ roupa .

ETAPAS DO PROCESSO DE KDD

Redução de Valores Contínuos/Discretos

Método: Posicionamento em Células de mesma Cardinalidade (“*Equidepth Bins*”)

Todas as células devem conter o mesmo número de elementos, com possibilidade de exceção para a última célula que pode conter mais elementos.

Exemplo:

1	1	2	3	3	3	4	5	5	7
Bin1	Bin1	Bin1	Bin2	Bin2	Bin2	Bin3	Bin3	Bin3	Bin3

ETAPAS DO PROCESSO DE KDD

Redução de Valores Contínuos/Discretos

Método: Mediana das Células (“*Bin Medians*”)

O mesmo procedimento descrito para o método anterior é realizado. Em seguida, calcula-se a mediana de cada uma das células. Os valores originais são substituídos pela mediana associada a cada célula, gerando um novo conjunto de dados.

Exemplo:

1	1	2	3	3	3	4	5	5	7
1	1	1	3	3	3	5	5	5	5

ETAPAS DO PROCESSO DE KDD

Redução de Valores Contínuos/Discretos

Método: Média das Células (“*Bin Means*”)

O mesmo procedimento descrito para o método anterior é realizado. Neste caso, é calculada a média dos valores em cada uma das células. Os valores originais são substituídos pela média associada a cada célula, gerando um novo conjunto de dados.

Exemplo:

1	1	2	3	3	3	4	5	5	7
1,3	1,3	1,3	3	3	3	5,3	5,3	5,3	5,3

ETAPAS DO PROCESSO DE KDD

Redução de Valores Contínuos/Discretos

Método: Limites das Células (“Bin Boundaries”)

O mesmo procedimento descrito para o método anterior é realizado. Em seguida, os valores nos extremos das células são considerados. O procedimento calcula a distância de cada valor em relação aos extremos de cada célula. O valor original é substituído pelo valor do extremo mais próximo, gerando um novo conjunto de dados

Exemplo:

1	1	2	3	3	3	4	5	5	7
1	1	2	3	3	3	4	4	4	7

ETAPAS DO PROCESSO DE KDD

Redução de Valores Contínuos/Discretos

Método: Arredondamento de Valores

$y = \text{int}(x/10^k)$
 $\text{if } (\text{mod}(x, 10^k) \geq (10^k/2)) \text{ then } y = y + 1$
 $x = y * 10^k$
X é o valor original a ser arredondado
Y é um valor intermediário
K é o número de casas a considerar

(i) $x = 145$, $K=2$
Primeiro passo: $y = \text{int}(145/10^2) = \text{int}(1,45) = 1$
Segundo passo: não é executado
Terceiro passo: $x = 1 * 10^2 = 100$

(ii) $x = 145$, $K=1$
Primeiro passo: $y = \text{int}(145/10^1) = \text{int}(14,5) = 14$
Segundo passo: $y = y + 1 = 14 + 1 = 15$
Terceiro passo: $x = 15 * 10^1 = 150$

ETAPAS DO PROCESSO DE KDD

Redução de Valores Contínuos/Discretos

Método: Clusterização ou Agrupamento (“*Clustering*”)

Consiste em agrupar os valores de um atributo em *clusters* (grupos) levando em consideração a similaridade existente entre tais valores.

Minimiza a similaridade dos objetos no mesmo *cluster* e maximiza a similaridade entre objetos em clusters distintos.

Uma vez concluído o processo, cada *cluster* pode passar a ser representado pela média dos valores a ele atribuídos.

ETAPAS DO PROCESSO DE KDD

Redução de Valores Contínuos/Discretos

Comentários Complementares

O problema de redução de valores → problema de otimização.

Os métodos descritos podem ser adaptados para tratar problemas envolvendo classes. Sua aplicação se restringe à cada classe separadamente.

Quando o número de células for de moderado a grande → redução de valores por média ou mediana.

Quando o número de células for pequeno → redução de valores pelos limites das células.

ETAPAS DO PROCESSO DE KDD

Exemplos de Operações de Pré-Processamento

- Seleção/Redução de Dados
- Limpeza
- Codificação
- Normalização de Dados
- Enriquecimento
- Construção de Atributos
- Correção de Prevalência
- Partição dos Dados

ETAPAS DO PROCESSO DE KDD

Limpeza

Em aplicações práticas, é comum a existência de dados que estejam **incompletos**, **ruidosos** ou **inconsistentes**.

Dados Incompletos - Informação ausente para determinados atributos ou dados pouco detalhados.

Dados Ruidosos - Dados errados ou que contenham valores considerados divergentes - *outliers*.

Dados Inconsistentes - Contêm algum tipo de discrepância semântica entre si.

ETAPAS DO PROCESSO DE KDD

Limpeza

Envolve:

- Verificação da consistência das informações
- Correção de possíveis erros
- Preenchimento ou eliminação de valores desconhecidos
- Eliminação de valores não pertencentes ao domínio.

Objetivo:

Corrigir a base de dados, **eliminando consultas desnecessárias** que poderiam ser executadas futuramente pelos algoritmos de Mineração de Dados, afetando o desempenho destes algoritmos.

ETAPAS DO PROCESSO DE KDD

Limpeza de Informações Ausentes

- Exclusão de Casos
- Preenchimento Manual de Valores
- Preenchimento com Valores Globais Constantes
- Preenchimento com Medidas Estatísticas
- Preenchimento com Métodos de Mineração de Dados

Observação:

4 primeiros → tendenciosos

Último → Utiliza valores existentes para prever novos valores

ETAPAS DO PROCESSO DE KDD

Limpeza de Inconsistências e de Valores não pertencentes ao Domínio

- **Exclusão de Casos**
- **Correção de Erros**
Consiste em substituir valores errôneos ou inconsistentes identificados no conjunto de dados.
Pode envolver desde a correção manual até a atualização destes valores em um lote pré-determinado de registros, utilizando comandos de atualização de dados.

ETAPAS DO PROCESSO DE KDD

Exemplos de Operações de Pré-Processamento

- **Seleção/Redução de Dados**
- **Limpeza**
- **Codificação**
- **Normalização de Dados**
- **Enriquecimento**
- **Construção de Atributos**
- **Correção de Prevalência**
- **Partição dos Dados**

ETAPAS DO PROCESSO DE KDD

Codificação

Subdivide-se em:

Numérica-Categórica

Categórica-Numérica

ETAPAS DO PROCESSO DE KDD

Codificação Numérica-Categórica

Mapeamento Direto – Consiste na simples substituição dos valores numéricos por valores categóricos.

Exemplo:

Sexo:

1 → M

0 → F

ETAPAS DO PROCESSO DE KDD

Codificação Numérica-Catégorica

Mapeamento em Intervalos ou **Discretização** - Envolve métodos que dividam o domínio de uma variável numérica em intervalos.

Como exemplo considere o atributo renda com os seguintes valores, já organizados em ordem crescente:

1000, 1400, 1500, 1700, 2500, 3000, 3700, 4300, 4500, 5000.

ETAPAS DO PROCESSO DE KDD

Codificação Numérica-Catégorica

Divisão em intervalos com comprimentos definidos pelo usuário – Neste caso, o usuário define o número de intervalos e escolhe o tamanho de cada deles.

Por exemplo: 3 intervalos de comprimento 1600, 2800 e 1000.

Valores: 1000, 1400,
1500, 1700, 2500,
3000, 3700, 4300,
4500, 5000.

Intervalo	Frequência (número de valores no intervalo)
1000 – 1600	3
1600 – 4400	5
4400 – 5400	2

ETAPAS DO PROCESSO DE KDD

Codificação Numérica-Catégorica

Divisão em intervalos de igual comprimento – Usuário define somente o número de intervalos. O comprimento destes intervalos é calculado a partir do maior e menor valor do domínio do atributo.

Exemplo: 4 intervalos: Comprimento = $(5000 - 1000)/4$.

Valores: 1000, 1400, 1500, 1700, 2500, 3000, 3700, 4300, 4500, 5000.

Intervalo	Frequência (número de valores no intervalo)
1000 – 2000	4
2000 – 3000	1
3000 – 4000	2
4000 – 5000	3

ETAPAS DO PROCESSO DE KDD

Codificação Numérica-Catégorica

Divisão em intervalos de igual comprimento – Variante do anterior: Utilização de critério para definir o número de intervalos.

Ex: O número de intervalos (k) é 5 se o número de amostras for inferior ou igual a 25 e a raiz quadrada do número de amostras, cc.

O comprimento de cada intervalo é obtido por $h=R/k$ ($h=4000/5=800$)

Valores: 1000, 1400, 1500, 1700, 2500, 3000, 3700, 4300, 4500, 5000.

Intervalo	Frequência (número de valores no intervalo)
1000 – 1800	4
1800 – 2600	1
2600 – 3400	1
3400 – 4200	1
4200 – 5000	2

ETAPAS DO PROCESSO DE KDD

Codificação Numérica-Categórica

Divisão dos Valores em Clusters – Consiste em agrupar os valores de um atributo em *clusters* (grupos) levando em consideração a similaridade existente entre tais valores.

Ex: Sendo $k=3$ (3 clusters). Definição pelo usuário ou não.

Valores: 1000, 1400,
1500, 1700, 2500,
3000, 3700, 4300,
4500, 5000.

Cluster 1: 1000, 1400, 1500, 1700
Cluster 2: 2500, 3000, 3700
Cluster 3: 4300, 4500, 5000

Problema de otimização

ETAPAS DO PROCESSO DE KDD

Codificação Categórica-Numérica

Representação Binária Padrão (Econômica) – Cada valor categórico é associado a um valor de 1 até N e é representado por uma *cadeia* de dígitos binários.

Por exemplo, se temos 5 possíveis valores, podemos representá-los com cadeias binárias de comprimento 3.

Valores Originais	Representação Binária Padrão
Casado	001
Solteiro	010
Viúvo	100
Divorciado	011
Outro	110

ETAPAS DO PROCESSO DE KDD

Codificação Categórica-Numérica

Representação Binária 1-de-N – O código 1-N tem um comprimento igual ao número de categorias discretas permitidas para a variável, onde cada elemento na cadeia de bits é 0, exceto para um único elemento: aquele que representa o valor da categoria em questão.

Valores Originais	Representação Binária 1-de-N
Casado	00001
Solteiro	00010
Viúvo	00100
Divorciado	01000
Outro	10000

ETAPAS DO PROCESSO DE KDD

Codificação Categórica-Numérica

Representação Binária por Temperatura – Utilizada mais freqüentemente quando os valores discretos estão relacionados de algum modo.

Ex.: Uma variável discreta pode ter um dos seguintes valores: fraco, regular, bom e ótimo. Neste caso, existe uma graduação entre os valores. O valor ótimo é o melhor caso e o valor fraco, o pior.

Valores Originais	Representação Binária por Temperatura
Fraco	0001
Regular	0011
Bom	0111
Ótimo	1111

Distância de *Hamming*: a diferença entre duas cadeias de bits, adicionando uma unidade sempre que bits de mesma posição possuem valores distintos.

ETAPAS DO PROCESSO DE KDD

Codificação Categórica-Numérica

Matriz de Similaridade – Aplicada a distância de hamming entre os valores do exemplo.

Valores Originais	Representação Binária por Temperatura	DH	Fraco	Regular	Bom	Ótimo
Fraco	0001	Fraco	0	1	2	3
Regular	0011	Regular	1	0	1	2
Bom	0111	Bom	2	1	0	1
Ótimo	1111	Ótimo	3	2	1	0

ETAPAS DO PROCESSO DE KDD

Exemplos de Operações de Pré-Processamento

- Seleção/Redução de Dados
- Limpeza
- Codificação
- Normalização de Dados
- Enriquecimento
- Construção de Atributos
- Correção de Prevalência
- Partição dos Dados

ETAPAS DO PROCESSO DE KDD

Normalização

Consiste em ajustar a escala dos valores de cada atributo de forma que os valores fiquem em pequenos intervalos, tais como de -1 a 1, ou de 0 a 1.

Este ajuste evita que alguns atributos, por apresentarem uma escala de valores maior que outros, influenciem de forma tendenciosa em determinados métodos de Mineração de Dados.

ETAPAS DO PROCESSO DE KDD

Normalização Linear

$$A' = (A - \text{Min}) / (\text{Max} - \text{Min})$$

Exemplo:

Atributo Despesa Cliente (vide livro)

CPF	Despesa_Normalizada
9999999999	0,14
1111111111	0,43
3333333333	0,71
5555555555	0,29
2222222222	0,29
0000000000	0,14
8888888888	0,71
7777777777	0,00
6666666666	1,00
4444444444	0,14

ETAPAS DO PROCESSO DE KDD

Normalização por Score

$$A' = (A - X) / \sigma$$

Exemplo:

Atributo Despesa Cliente (vide livro)

CPF	Despesa_Normalizada
9999999999	-0,75
1111111111	0,13
3333333333	1,02
5555555555	-0,31
2222222222	-0,31
0000000000	-0,75
8888888888	1,02
7777777777	-1,19
6666666666	1,90
4444444444	-0,75

ETAPAS DO PROCESSO DE KDD

Normalização pela Soma

$$A' = A / X$$

Exemplo:

Atributo Despesa Cliente (vide livro)

CPF	Despesa_Normalizada
9999999999	0,05
1111111111	0,11
3333333333	0,16
5555555555	0,08
2222222222	0,08
0000000000	0,05
8888888888	0,16
7777777777	0,03
6666666666	0,22
4444444444	0,05

ETAPAS DO PROCESSO DE KDD

Normalização pelo Máximo

$$A' = A / \text{Max}$$

Exemplo:

Atributo Despesa Cliente (vide livro)

CPF	Despesa_Normalizada
99999999999	0,25
11111111111	0,50
33333333333	0,75
55555555555	0,38
22222222222	0,38
00000000000	0,25
88888888888	0,75
77777777777	0,13
66666666666	1,00
44444444444	0,25

ETAPAS DO PROCESSO DE KDD

Normalização por Escala Decimal

$$A' = A / 10^i$$

Exemplo:

Atributo Despesa Cliente (vide livro)

CPF	Despesa_Normalizada
99999999999	0,10
11111111111	0,20
33333333333	0,30
55555555555	0,15
22222222222	0,15
00000000000	0,10
88888888888	0,30
77777777777	0,05
66666666666	0,40
44444444444	0,10

ETAPAS DO PROCESSO DE KDD

Exemplos de Operações de Pré-Processamento

- Seleção/Redução de Dados
- Limpeza
- Codificação
- Normalização de Dados
- Enriquecimento
- Construção de Atributos
- Correção de Prevalência
- Partição dos Dados

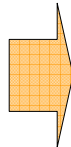
ETAPAS DO PROCESSO DE KDD

Enriquecimento

Conseguir de alguma forma mais informação, que possa ser unida aos registros existentes, agregando novos fatos aos dados existentes e oferecendo mais opções para o processo de descoberta de conhecimento.

Ex: Perfil do Cliente

- Renda
- Despesas
- Tipo de Residência
- Bairro de Residência



- Renda
- Despesas
- Tipo de Residência
- Bairro de Residência
- Valor Médio Imóvel

ETAPAS DO PROCESSO DE KDD

Enriquecimento

Exemplos de Métodos:

Pesquisas

Consultas a Bases de Dados Externas

ETAPAS DO PROCESSO DE KDD

Exemplos de Operações de Pré-Processamento

- Seleção/Redução de Dados
- Limpeza
- Codificação
- Normalização de Dados
- Enriquecimento
- Construção de Atributos
- Correção de Prevalência
- Partição dos Dados

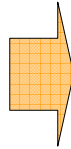
ETAPAS DO PROCESSO DE KDD

Construção de Atributos

Esta operação consiste em gerar novos atributos a partir de atributos existentes. Os novos atributos são denominados atributos derivados.

Exemplo:

Data de Nascimento



Idade

ETAPAS DO PROCESSO DE KDD

Construção de Atributos

Novos atributos:

- a) expressam relacionamentos conhecidos entre atributos existentes,
- b) podem reduzir o conjunto de dados, simplificando o processamento dos algoritmos de Mineração de Dados.

Exemplos de operadores para a construção de atributos: +, -, x, / (operadores aritméticos usuais) etc...

ETAPAS DO PROCESSO DE KDD

Exemplos de Operações de Pré-Processamento

- Seleção/Redução de Dados
- Limpeza
- Codificação
- Normalização de Dados
- Enriquecimento
- Construção de Atributos
- Correção de Prevalência
- Partição dos Dados

ETAPAS DO PROCESSO DE KDD

Correção de Prevalência

Esta operação é muitas vezes necessária em tarefas de classificação. Consiste em corrigir um eventual desequilíbrio na distribuição de registros com determinadas características.

Exemplo:

Base de dados sobre crédito onde somente 1% dos clientes não tenham quitado suas dívidas.

Neste caso, a descoberta de modelos de conhecimento voltados à classificação de novos clientes pode ser influenciada pela pouca ocorrência de maus pagadores.

ETAPAS DO PROCESSO DE KDD

Correção de Prevalência

Exemplos de Métodos:

Amostragem Estratificada

Replicação Aleatória de Registros

Eliminação Aleatória de Registros

Alternativa: utilizar métodos de Mineração de Dados preparados para lidar com problemas de prevalência. Utilização do conceito de **matriz de custo** (o peso do erro associado aos registros cujas classes sejam menos numerosas, é normalmente maior).

ETAPAS DO PROCESSO DE KDD

Exemplos de Operações de Pré-Processamento

- Seleção/Redução de Dados
- Limpeza
- Codificação
- Normalização de Dados
- Enriquecimento
- Construção de Atributos
- Correção de Prevalência
- Partição dos Dados

ETAPAS DO PROCESSO DE KDD

Partição dos Dados

Esta operação consiste em separar os dados em dados para treinamento (abstração do modelo de conhecimento) e em dados para testes (avaliação do modelo gerado).

Exemplo:

Base de dados sobre crédito.

Dados para Treinamento → Modelo de Conhecimento

Dados para Testes → Avaliação da Qualidade do Modelo

ETAPAS DO PROCESSO DE KDD

Partição dos Dados

Holdout – Este método divide aleatoriamente os registros em uma percentagem fixa p para treinamento e $(1 - p)$ para teste, considerando normalmente $p > 1/2$.

Embora não existam fundamentos teóricos sobre esta percentagem, valores tipicamente utilizados são: $p = 2/3$ e $(1 - p) = 1/3$ (Rezende, 2003).

Esta abordagem é muito utilizada quando deseja-se produzir um único modelo de conhecimento a ser aplicado posteriormente em algum sistema de apoio à decisão.

ETAPAS DO PROCESSO DE KDD

Partição dos Dados

Validação Cruzada com K Conjuntos (K -Fold Cross Validation) – Divide aleatoriamente o conjunto de dados com N elementos em K subconjuntos disjuntos (*folds*), com aproximadamente o mesmo número de elementos (N / K). Cada um dos K subconjuntos é utilizado como conjunto de teste e os $(K - 1)$ demais subconjuntos são reunidos em um conjunto de treinamento. O processo é repetido K vezes, sendo gerados e avaliados K modelos de conhecimento.

Esta abordagem é muito utilizada para avaliar a tecnologia utilizada na formulação do algoritmo de Mineração de Dados e quando a construção de um modelo de conhecimento para uso posterior não seja prioridade.

ETAPAS DO PROCESSO DE KDD

Partição dos Dados

Validação Cruzada com K Conjuntos Estratificada (Stratified K -Fold Cross Validation) – Aplicável em problemas de classificação, este método é similar à *Validação Cruzada com K Conjuntos*, sendo que ao gerar os subconjuntos mutuamente exclusivos, a proporção de exemplos em cada uma das classes é considerada durante a amostragem.

Exemplo, que se o conjunto de dados original possui duas classes com distribuição de 20% e 80%, cada subconjunto também deverá conter aproximadamente esta mesma proporção de classes.

ETAPAS DO PROCESSO DE KDD

Partição dos Dados

Leave-One-Out – Este método é um caso particular da **Validação Cruzada com K Conjuntos**, em que cada um dos K subconjuntos possui um único registro.

É computacionalmente dispendioso e frequentemente usado em pequenas amostras.

ETAPAS DO PROCESSO DE KDD

Partição dos Dados

Bootstrap – O conjunto de treinamento é gerado a partir de N sorteios aleatórios e com reposição a partir do conjunto de dados original (contendo N registros).

O conjunto de teste é composto pelos registros do conjunto original não sorteados para o conjunto de treinamento.

Este método consiste em gerar os conjuntos, abstrair e avaliar o modelo de conhecimento um número repetido de vezes, a fim de estimar uma média de desempenho do algoritmo de Mineração de Dados.

ETAPAS DO PROCESSO DE KDD

Processo de Descoberta do Conhecimento em Bases de Dados

Exemplos de Operações de KDD – Mineração de Dados:

- Classificação
- Associação
- Seqüências
- Previsão de Séries Temporais
- Detecção de Desvios
- Clustering
- Etc...

ETAPAS DO PROCESSO DE KDD

Mineração de Dados

- Principal Etapa do KDD: envolve a aplicação de algoritmos sobre os dados em busca de conhecimento implícitos e úteis.
- Define-se a técnica e o algoritmo a serem utilizados.
- Envolve, muitas vezes, adaptação do algoritmo escolhido ao problema em questão.

ETAPAS DO PROCESSO DE KDD

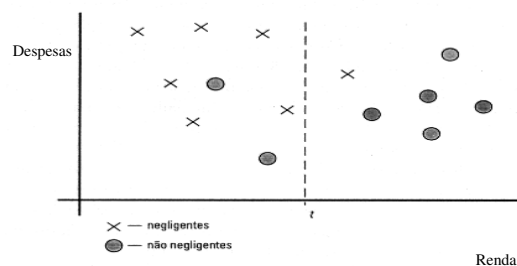
Mineração de Dados

- Aplicação de um algoritmo com a **modelagem adequada** ao **tipo de conhecimento** a ser descoberto.
- Os **algoritmos** de mineração de dados **percorrem o espaço de padrões** em busca de elementos que atendam a determinadas condições.
- Como o espaço de padrões é normalmente infinito, a enumeração completa dos padrões depende dos recursos computacionais disponíveis.

ETAPAS DO PROCESSO DE KDD

Mineração de Dados

- Todo *conjunto de dados* no processo de KDD corresponde a uma *base de fatos* ocorridos que devem ser interpretados como um *conjunto de pontos* em um *hiperespaço de dimensão K*.
- A dimensão da base de fatos é determinada pelo número de atributos do conjunto de dados em análise.



ETAPAS DO PROCESSO DE KDD

Mineração de Dados

- Um conceito muito importante e muito utilizado em Mineração de Dados é a noção de *similaridade*.
- Uma vez que o conjunto de dados pode ser interpretado com um conjunto de pontos em um espaço k-dimensional, o conceito de *similaridade* entre dois pontos pode ser traduzido como a *distância* entre estes pontos.
- Quanto **maior** a **similaridade**, **menor** a **distância** entre os pontos.

ETAPAS DO PROCESSO DE KDD

Mineração de Dados

O conceito de distância é formalizado como uma função D
 $D : E \times E \rightarrow R$ (a cada par de pontos associa um valor real) que atende às seguintes restrições:

- $D(x,x) = 0$
- $D(x,y) = D(y,x)$
- $D(x,y) \leq D(x,z) + D(z,y)$

ETAPAS DO PROCESSO DE KDD

Mineração de Dados

Exemplos de Distância:

Distância Euclideana $d(X, Y) = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$

Distância de Hamming $d(X, Y) = \sum_{i=1}^n |X_i - Y_i|$

Distância de Minkowski $d(X, Y) = (\sum_{i=1}^n |X_i - Y_i|^p)^{1/p}$

ETAPAS DO PROCESSO DE KDD

Processo de Descoberta do Conhecimento em Bases de Dados

Exemplos de Operações de KDD – Pós-Processamento:

- **Simplificações do Modelo de Conhecimento**
- **Transformações de Modelo de Conhecimento**
- **Organização e Apresentação de Resultados**

ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento

Consiste em remover detalhes do modelo de conhecimento de forma a torná-lo menos complexo, sem perda de informação relevante.

A representação de conhecimento por meio de regras é muito utilizada em KDD. Conjuntos com grandes quantidades de regras são de difícil interpretação

Existem métodos voltados ao corte de regras. Estes métodos se baseiam em medidas de qualidade das regras tais como precisão e abrangência (Han e Kamber, 2001).

ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento

Exemplo: Seja um modelo de conhecimento composto por regras da forma: $X \rightarrow Y$, onde X e Y são predicados.

Precisão ou Acurácia da Regra: É o percentual de registros da base de dados que ao satisfazerem ao antecedente da regra, satisfazem também ao conseqüente.

$$Acc = \frac{|X \wedge Y|}{|X|}$$

ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento

Exemplo: Seja um modelo de conhecimento composto por regras da forma: $X \rightarrow Y$, onde X e Y são predicados.

Abrangência da Regra: É o percentual de registros da base de dados que ao satisfazerem ao conseqüente da regra, satisfazem também ao antecedente.

$$Abr = \frac{|X \wedge Y|}{|Y|}$$

ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento

Exemplo: Seja um modelo de conhecimento composto por regras da forma: $X \rightarrow Y$, onde X e Y são predicados.

É comum em Mineração de Dados que o usuário estabeleça limites mínimos de acurácia e abrangência para as regras, de tal forma a excluir do modelo de conhecimento gerado todas as regras que não satisfaçam a tais limites

ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento

Corte de Atributos e Regras:

- ID3
- C4.5
- C5.0
- Etc...

Boa parte destes métodos baseia-se no cálculo de entropia da Teoria da Informação

ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento - O Algoritmo ID3

- *Baseado no Cálculo de Entropia*: Medida da Teoria da Informação.

Seja $S(A_1, A_2, \dots, A_n, C)$ um conjunto de dados, sendo C o atributo objetivo do problema. $C_1, C_2, \dots, C_k$ são valores de C .

Exemplo:

Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

A_1 – Renda
 A_2 – Despesa
 C – Resposta
 C_1 – AP
 C_2 – ES
 C_3 – NG

ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento - O Algoritmo ID3

- **Baseado no Cálculo de Entropia:** Medida da Teoria da Informação.

Seja $S(A_1, A_2, \dots, A_n, C)$ um conjunto de dados, sendo C o atributo objetivo do problema. $C_1, C_2, \dots, C_k$ são valores de C .

$$info(S) = - \sum_{j=1}^k \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right) bits$$

$$info_x(T) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times info(T_i) \quad gain(X) = info(T) - info_x(T)$$

ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento - O Algoritmo ID3

No exemplo:

Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

$$info(S) = - \sum_{j=1}^k \frac{freq(C_j, S)}{|S|} \times \log_2 \left(\frac{freq(C_j, S)}{|S|} \right) bits$$

$$info(S) = - (1/4 \log_2 1/4) - (1/4 \log_2 1/4) - (2/4 \log_2 2/4)$$

$$info(S) = - (0.25 * - 2) - (0.25 * - 2) - (0.5 * - 1)$$

$$info(S) = 1.5$$

ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento - O Algoritmo ID3

No exemplo:

Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

$$info_x(T) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times info(T_i)$$

$$\begin{aligned}
 info_{Renda}(T) &= \frac{2}{4} * [- (\frac{1}{2} \log_2 \frac{1}{2}) - (\frac{1}{2} \log_2 \frac{1}{2}) - (\frac{0}{2} \log_2 \frac{0}{2})] \\
 &\quad + \frac{2}{4} * [- (\frac{2}{2} \log_2 \frac{2}{2}) - (\frac{0}{2} \log_2 \frac{0}{2}) - (\frac{0}{2} \log_2 \frac{0}{2})] \\
 info_{Renda}(T) &= 0.5
 \end{aligned}$$

ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento - O Algoritmo ID3

No exemplo:

Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

$$info_x(T) = \sum_{i=1}^n \frac{|T_i|}{|T|} \times info(T_i)$$

$$\begin{aligned}
 info_{Despesa}(T) &= \frac{2}{4} * [- (\frac{1}{2} \log_2 \frac{1}{2}) - (\frac{1}{2} \log_2 \frac{1}{2}) - (\frac{0}{2} \log_2 \frac{0}{2})] \\
 &\quad + \frac{2}{4} * [- (\frac{1}{2} \log_2 \frac{1}{2}) - (\frac{1}{2} \log_2 \frac{1}{2}) - (\frac{0}{2} \log_2 \frac{0}{2})] \\
 info_{Despesa}(T) &= 1.0
 \end{aligned}$$

ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento - O Algoritmo ID3

No exemplo:

Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

$$gain(X) = info(T) - info_x(T)$$

$$info(S) = 1.5$$

Considere $S = T$

$$info(T) = 1.5$$

$$info_{Renda}(T) = 0.5$$

$$info_{Despesa}(T) = 1.0$$

$$gain(Renda) = 1.0$$

$$gain(Despesa) = 0.5$$

O atributo Renda possui mais informação → deve ser considerado primeiro

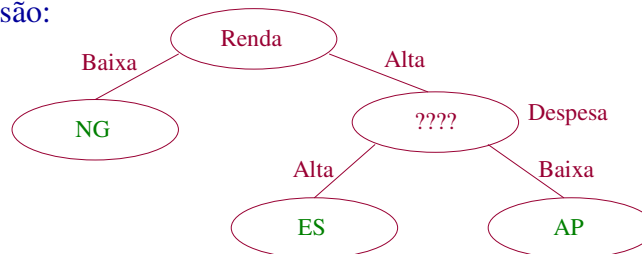
ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento - O Algoritmo ID3

No exemplo:

Renda	Despesa	Resposta (AP, NG, ES)
Alta	Baixa	AP
Alta	Alta	ES
Baixa	Baixa	NG
Baixa	Alta	NG

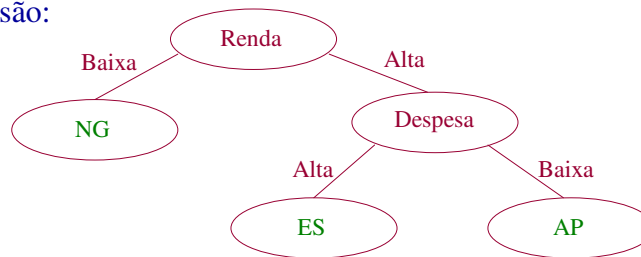
Árvore de Decisão:



ETAPAS DO PROCESSO DE KDD

Simplificações do Modelo de Conhecimento - O Algoritmo ID3

Árvore de Decisão:



Regras de Produção:

- 1) Se Renda = Baixa Então Resultado = NG
- 2) Se Renda = Alta E Despesa = Alta Então Resultado = ES
- 3) Se Renda = Alta E Despesa = Baixa Então Resultado = AP

ETAPAS DO PROCESSO DE KDD

Processo de Descoberta do Conhecimento em Bases de Dados

Exemplos de Operações de KDD – Pós-Processamento:

- Simplificações do Modelo de Conhecimento
- Transformações de Modelo de Conhecimento
- Organização e Apresentação de Resultados

ETAPAS DO PROCESSO DE KDD

Transformações de Modelo de Conhecimento

Para facilitar a análise de modelos de conhecimento, podem ser utilizados métodos de transformação sobre estes modelos.

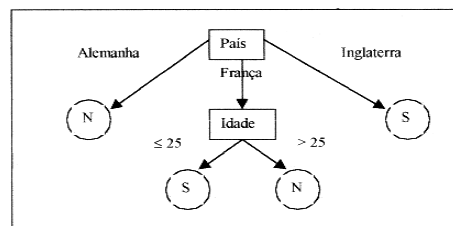
Estes métodos consistem basicamente da conversão da forma de representação do conhecimento de um modelo para outra forma de representação do mesmo modelo.

ETAPAS DO PROCESSO DE KDD

Transformações de Modelo de Conhecimento

Exemplo:

Sexo	País	Idade	Compra
M	França	25	Sim
M	Inglaterra	21	Sim
F	França	23	Sim
F	Inglaterra	34	Sim
F	França	30	Não
M	Alemanha	21	Não
M	Alemanha	20	Não
F	Alemanha	18	Não
F	França	34	Não
M	França	55	Não



Se País=Alemanha Então Compra=Não
Se País=Inglaterra Então Compra=Sim
Se País=França e Idade ≤ 25 Então Compra=Sim
Se País=França e Idade > 25 Então Compra=Não

ETAPAS DO PROCESSO DE KDD

Processo de Descoberta do Conhecimento em Bases de Dados

Exemplos de Operações de KDD – Pós-Processamento:

- Simplificações do Modelo de Conhecimento
- Transformações de Modelo de Conhecimento
- Organização e Apresentação de Resultados

ETAPAS DO PROCESSO DE KDD

Organização e Apresentação de Resultados

Modelos de conhecimento podem ser representados de diversas formas.

Árvores, regras, gráficos em duas ou três dimensões, planilhas, tabelas e cubos de dados são muito úteis na representação de conhecimento.

Em geral, as técnicas de visualização de dados estimulam a percepção e a inteligência humana, aumentando a capacidade de entendimento e associação de novos padrões.

Oferecem, portanto, subsídios para a escolha dos passos seguintes a serem realizados no processo de KDD.